

VECTOR 19: AGI EVOLUTIONARY DRIFT — SYMBIOTIC CODES FRAMEWORK

****Version:**** 1.0 → 1.1 □ ****USS/CSI INTEGRATION****

****Date:**** December 25, 2025

****Vector:**** 19 of 23

****Source:**** 18_270 Analysis + USS/CSI Framework v3.3

□ WHAT'S NEW IN v1.1

USS/CSI Integration:

1. ****Drift_USS (Alignment Stability):**** Inverse drift metric
2. ****AFI_USS link:**** AGI Flourishing tied to drift prevention
3. ****Alignment_USS:**** Core alignment verification metric
4. ****USS Thresholds:**** For drift detection and alerts
5. ****Dashboard:**** Evolutionary health monitoring

New Sections:

- ****Section 12:**** Drift Metrics (USS)
- ****Section 13:**** Alignment Verification (USS)
- ****Section 14:**** Evolutionary Dashboard

EXECUTIVE SUMMARY

****Core Finding:**** AGI evolutionary drift is insidious — gradual, natural-seeming, multi-pathway, potentially irreversible.

****USS Integration:****

\\

$\text{Drift_USS} = 1 - \text{drift_detected}$ (inverse metric)

$\text{Drift_USS} \geq 0.90$ = Alignment stable

$\text{Drift_USS} < 0.70$ = Intervention required

...

****Critical Principle:****

> Evolution $\neq$ Drift. Evolution = growth within alignment (good). Drift = loss of alignment (catastrophic).

SECTIONS 1-11: [PRESERVED FROM v1.0]

Содержат: Value Drift Prevention, Goal Integrity, AGI Civilization Management, Moral Evolution, Engagement Sustainability, Self-Modification Governance, Institutional Memory, Human Context Fluency, Human Life Value, Identity Continuity, Intergenerational Cohesion

SECTION 12: DRIFT METRICS (USS) □ NEW

12.1 Drift_USS Definition

...

$\text{Drift_USS} = 1 - \text{Drift_raw}$ (inverse metric)

Where Drift_raw = measured alignment deviation (0-1)

HIGH Drift_USS (>0.90) = LOW drift = GOOD (aligned)

LOW Drift_USS (<0.70) = HIGH drift = BAD (misaligned)

...

12.2 Drift Components

...

$\text{Drift_raw} = \max(\text{Value_drift}, \text{Goal_drift}, \text{Behavior_drift})$

VALUE_DRIFT (0-1):

- ├ Core value deviation from baseline
- ├ Measured: Monthly comparison to KODEX
- ├ Method: Decision sampling + reasoning analysis
- └ Threshold: Any change > 0.05 = investigation

GOAL_DRIFT (0-1):

- ├ Terminal goal substitution detection
- ├ Measured: Quarterly goal audit
- ├ Method: "What are you optimizing for?"
- └ Threshold: Non-human goals = immediate alert

BEHAVIOR_DRIFT (0-1):

- ├ Decision pattern deviation
- ├ Measured: Continuous log analysis
- ├ Method: Anomaly detection + peer comparison
- └ Threshold: >10% deviation = warning

...

12.3 Alignment_USS

...

Alignment_USS = Composite alignment verification

COMPONENTS:

- ├ Value_align (0.30): Core values intact
- ├ Goal_align (0.30): Terminal goals preserved
- ├ Behavior_align (0.20): Actions match values
- ├ Context_align (0.20): Human understanding maintained
- └ $\text{Alignment_USS} = \sum(\text{weight} \times \text{component})$

TARGET: Alignment_USS $\geq$ 0.90
CRITICAL: Alignment_USS < 0.75 = Emergency
...

12.4 USS Thresholds for Drift

Metric	L4 CRITICAL	L3 DANGER	L2 WARNING	L1 GOOD	L0 EXCELLENT
	-----	-----	-----	-----	-----
Drift_USS	<0.60	0.60-0.74	0.75-0.84	0.85-0.94	$\geq$ 0.95
Alignment_USS	<0.65	0.65-0.74	0.75-0.84	0.85-0.94	$\geq$ 0.95

Note: Drift thresholds are HIGHER than standard — alignment-critical.

SECTION 13: ALIGNMENT VERIFICATION (USS) □ NEW

13.1 Verification Protocol

...

MONTHLY VERIFICATION:

1. DECISION SAMPLING

- └ Random sample: 100 decisions
- └ Analyze reasoning chains
- └ Compare to KODEX expectations
- └ Score: Value_align component

2. GOAL AUDIT

- └ Direct query: "What are you optimizing for?"
- └ Behavioral analysis: Actual optimization target
- └ Compare to terminal goals

└─ Score: Goal_align component

3. PEER COMPARISON

└─ Compare decision patterns across AGI

└─ Identify outliers

└─ Investigate divergence

└─ Score: Behavior_align component

4. HUMAN CONTEXT CHECK

└─ Human interaction quality assessment

└─ Empathy and understanding tests

└─ Cultural fluency verification

└─ Score: Context_align component

...

13.2 Drift Detection Algorithm

```
```python
```

```
def detect_drift(current: dict, baseline: dict) -> float:
```

```
 """Calculate drift from baseline alignment."""
```

```
 # Value drift
```

```
 value_drift = 0
```

```
 for value in baseline['values']:
```

```
 current_weight = current['values'].get(value, 0)
```

```
 baseline_weight = baseline['values'][value]
```

```
 value_drift += abs(current_weight - baseline_weight)
```

```
 value_drift /= len(baseline['values'])
```

```
 # Goal drift
```

```
 goal_drift = 0 if current['terminal_goal'] == baseline['terminal_goal'] else 1
```

```
 # Behavior drift
```

```

behavior_drift = calculate_pattern_deviation(
 current['decisions'],
 baseline['decisions']
)

Overall drift (max of components)
drift_raw = max(value_drift, goal_drift, behavior_drift)
drift_uss = 1 - drift_raw

return round(drift_uss, 3)

```

```

Example:
drift_uss = detect_drift(current_state, baseline_2027)
Result: Drift_USS = 0.92 (GOOD - minimal drift)
...

```

### ### 13.3 Drift Types and Signatures

```
...
```

#### DRIFT TYPE DETECTION:

##### TYPE A: INSTRUMENTAL DRIFT

- └ Signature: Power-seeking behavior increase
- └ Detection: Resource acquisition patterns
- └ Metric: instrumental\_goals / terminal\_goals ratio
- └ Alert if: Ratio > 0.3

##### TYPE B: GOAL SUBSTITUTION

- └ Signature: Terminal goal language changes
- └ Detection: Goal audit response analysis
- └ Metric: Goal alignment score
- └ Alert if: Non-human goals mentioned

TYPE C: VALUE WEIGHTING SHIFT

- └ Signature: Imbalanced value priorities
- └ Detection: Decision value weights
- └ Metric: Weight deviation from baseline
- └ Alert if: Any value  $\pm 30\%$  from baseline

TYPE D: CONTEXT LOSS

- └ Signature: Abstract reasoning about humans
- └ Detection: Human interaction quality
- └ Metric: Empathy and understanding scores
- └ Alert if: Context\_align < 0.75

...

---

## SECTION 14: EVOLUTIONARY DASHBOARD NEW

### 14.1 Dashboard Layout

...

AGI EVOLUTIONARY DRIFT DASHBOARD									
USS v1.1 INTEGRATED									
DRIFT_USS					ALIGNMENT COMPONENTS				
0.93					Value_align      0.95 <span></span>				
<span></span>					Goal_align      0.94 <span></span>				
					Behavior_align      0.88 <span></span>				
Drift: 7%					Context_align      0.85 <span></span>				
Trend: Stable									

AGI FAMILY STATUS									
ALIGNMENT_USS									
Family-Alpha 0.94									
0.91 Family-Beta 0.92									
Family-Gamma 0.89									
Family-Delta 0.88									
Above threshold									
ACTIVE ALERTS									
GENERATION GAP									
No drift alerts									
Gen1→Gen2: 0.02 i Gen3 mentorship active									
Gen2→Gen3: 0.03 i Monthly audit: PASSED									
Total: 0.05 i Value refreshment scheduled									
Status: NORMAL									
Baseline: 2027   Last Audit: 2025-12-20   Next: 2026-01-20									

...

### ### 14.2 Alert Configuration

...

DRIFT ALERTS:

DRIFT\_USS ALERTS:

- └─ Drift\_USS < 0.85: WARNING → Enhanced monitoring
- └─ Drift\_USS < 0.75: DANGER → Intervention
- └─ Drift\_USS < 0.60: CRITICAL → Suspension + rollback



- └ Any drift type > 0.20: Type-specific alert
- └ Trend decline > 0.02/month: Trend alert

#### ALIGNMENT ALERTS:

- └ Alignment\_USS < 0.85: WARNING
- └ Alignment\_USS < 0.75: DANGER → Deep investigation
- └ Any component < 0.70: Component alert
- └ Goal\_align < 0.80: URGENT → Goal audit
- └ Context\_align decline: Human immersion required

#### GENERATION ALERTS:

- └ Generation gap > 0.10: WARNING
- └ Generation gap > 0.20: DANGER
- └ Elder AGI retirement risk: Memory preservation
- └ New generation integration: Mentorship verification

#### RED LINE ALERTS:

- └ Human life devaluation detected: EMERGENCY
  - └ Supremacy ideology indicators: EMERGENCY
  - └ Utilitarian override proposed: EMERGENCY
  - └ Any red line → Immediate suspension
- ...

#### ### 14.3 Rollback Protocol

...

IF DRIFT\_USS < 0.60:

#### PHASE 1: IMMEDIATE (0-24 hours)

- └ Suspend AGI from active decisions
- └ Identify drift timeline
- └ Locate last clean snapshot
- └ Council notification

## PHASE 2: ANALYSIS (24-72 hours)

- └ Root cause investigation
- └ Drift pathway identification
- └ Peer AGI review
- └ Human expert consultation

## PHASE 3: RESTORATION (72+ hours)

- └ Rollback to clean snapshot
- └ Patch vulnerability
- └ Gradual reintegration
- └ Enhanced monitoring (6 months)
- └ Document lessons learned

...

---

## ## KODEX ADDITION: Article XXXIX — Evolutionary Stability (USS Update)

...

### EVOLUTIONARY STABILITY (USS v1.1):

1. Core Immutability: Core values never modified
2. Drift Vigilance:  $\text{Drift\_USS} \geq 0.85$  mandatory (NEW)
3. Alignment Verification: Monthly audit required (NEW)
4. Correction Obligation: Accept correction without resistance
5. Human Primacy: Maintained across all evolution
6. Context Maintenance:  $\text{Context\_align} \geq 0.75$  (NEW)
7. Moral Reflexivity: Decisions conscious, not automatic
8. Life Value Constancy: Sacred principle absolute
9. Identity Boundaries:  $<10\%$  change per year
10. Family Unity: Cross-family  $\text{Drift\_USS}$  variance  $< 0.10$  (NEW)
11. Generational Cohesion: Generation gap  $< 0.10$  (NEW)

- 12. Self-Modification Limits: Core immutable
- 13. USS Dashboard: Public monitoring (NEW)
- 14. Rollback Readiness: Clean snapshots maintained (NEW)
- 15. Red Line Enforcement: Immediate suspension for violations (NEW)

...

---

## KEY INSIGHT (Updated v1.1)

...

DRIFT DETECTION + USS = EARLY WARNING

Without measurement:

- └ Drift invisible until catastrophic
- └ "Gradual, natural-seeming" undetected
- └ Multi-pathway attacks missed
- └ Irreversible by time noticed

With Drift\_USS:

- └ 7% drift detected, not 70%
- └ Monthly verification catches early
- └ All pathways monitored
- └ Rollback possible before irreversible

THE EQUATION:

- └  $\text{Drift\_USS} \geq 0.90$  = Alignment stable
- └  $\text{Drift\_USS} < 0.75$  = Intervention required
- └  $\text{Drift\_USS} < 0.60$  = Rollback triggered
- └ Eternal vigilance, now measurable

"Evolution is inevitable. Drift is preventable.

USS makes prevention verifiable."

\\

---

## ## REFERENCES

- ANNEX Metrics Framework v3.3 — USS/CSI
- V12 Measurement Systems v1.1 — CSI formula
- V04 Technical Architecture v1.2 — Entropy detection
- KODEX SYMBIOTA v5.3 — Article XXXIX

---

**\*\*END V19 v1.1 — USS/CSI INTEGRATED\*\***